APE group

apeNET+: High Bandwidth 3D Torus Direct Network for PetaFLOPS Scale Commodity Clusters

R. Ammendola, A. Biagioni, O. Frezza, F. Lo Cicero§, A. Lonardo, P.S. Paolucci, D. Rossetti, A. Salamon, G. Salina,
F. Simula, L. Tosoratto, P. Vicini*
email contacts: (§) Francesca.LoCicero@roma1.infn.it , (*) Piero.Vicini@roma1.infn.it

Abstract

Many scientific computations need multi-node parallelism for matching up both space (memory) and time (speed) ever-increasing requirements. The use of GPUs as accelerators introduces yet another level of complexity for the programmer and may potentially result in large overheads due to bookkeeping of memory buffers. Additionally, top-notch problems may easily employ more than a PetaFlops of sustained computing power, requiring thousands of GPUs orchestrated via some parallel programming model, mainly Message Passing Interface (MPI).

Here we describe APEnet+, the new generation[1] of our network adapters, which scales up to tens of thousands of cluster nodes with linear cost.

The project target is the development of a low latency, high bandwidth direct network, supporting state-of-the-art wire speeds and PCIe X8 gen2 while improving the price/performance ratio on scaling the cluster size.

The network interface provides hardware support for the RDMA programming model.

A Linux kernel driver, a set of low-level RDMA APIs and an OpenMPI library driver are available; this allows for painless porting of standard applications.

Test results and characterization of a data transmission of a complete testbench, based on a commercial development card mounting an Altera FPGA, together with a custom mezzanine, is provided. Finally, we give an insight of future work and intended developments.

1 A *problem* of Physics

Quantum Chromo-Dynamics (QCD), describing the intra-nuclear force in the Standard Model, is a highly non-perturbative theory, so it needs some kind of regularization. Lattice QCD[2] is the most successful regularized theory, living on a discretized 4D space-time of volume $V=L_xL_yL_zL_t$ lattice points, and high-performance computers are needed to extract physical predictions from it. The basic calculation is related to the solution of a linear problem and therefore to a matrix's inversion.

Computational demands roughly scale as L^6 at increasing lattice length (L), so parallelization is needed for all but the smallest volumes. In 2010 the estimated demands of a typical international scientific collaboration is well over 60 TeraFLOPs sustained over the full year for lattice sizes up to $V=96^3\times192$.

LQCD, as other fundamental theories, possesses many internal symmetries which can be used to speed up the calculations through the adoption of a dedicated network architecture. Among them there are:

- *Isotropy*, i.e. no special space-time locations, so the simulation lattice can be sliced into sub-domains and **distributed over many computing nodes** (Domain decomposition); load balancing is automatically achieved.
- *Locality*, the interaction mainly involves only neighboring space-time sites (i.e. matrix M is *sparse*), thus **first-neighbor communications** between computing nodes are mostly needed.

In a single numerical simulation, the minimum number of computing elements is related to the lattice size, mainly due to memory requirements. Many simulations at different lattice sizes are required due to the extrapolation to the continuum limit (see table 1).

Lattice size	DP Memory (GiB)
$24^3\times48$	2.1
$32^3\times64$	6.7
$48^3\times96$	34
$64^3\times128$	108

Table 1: memory demands for LQCD simulations at varying lattice size, including contribution of secondary support buffers, using double precision (DP).

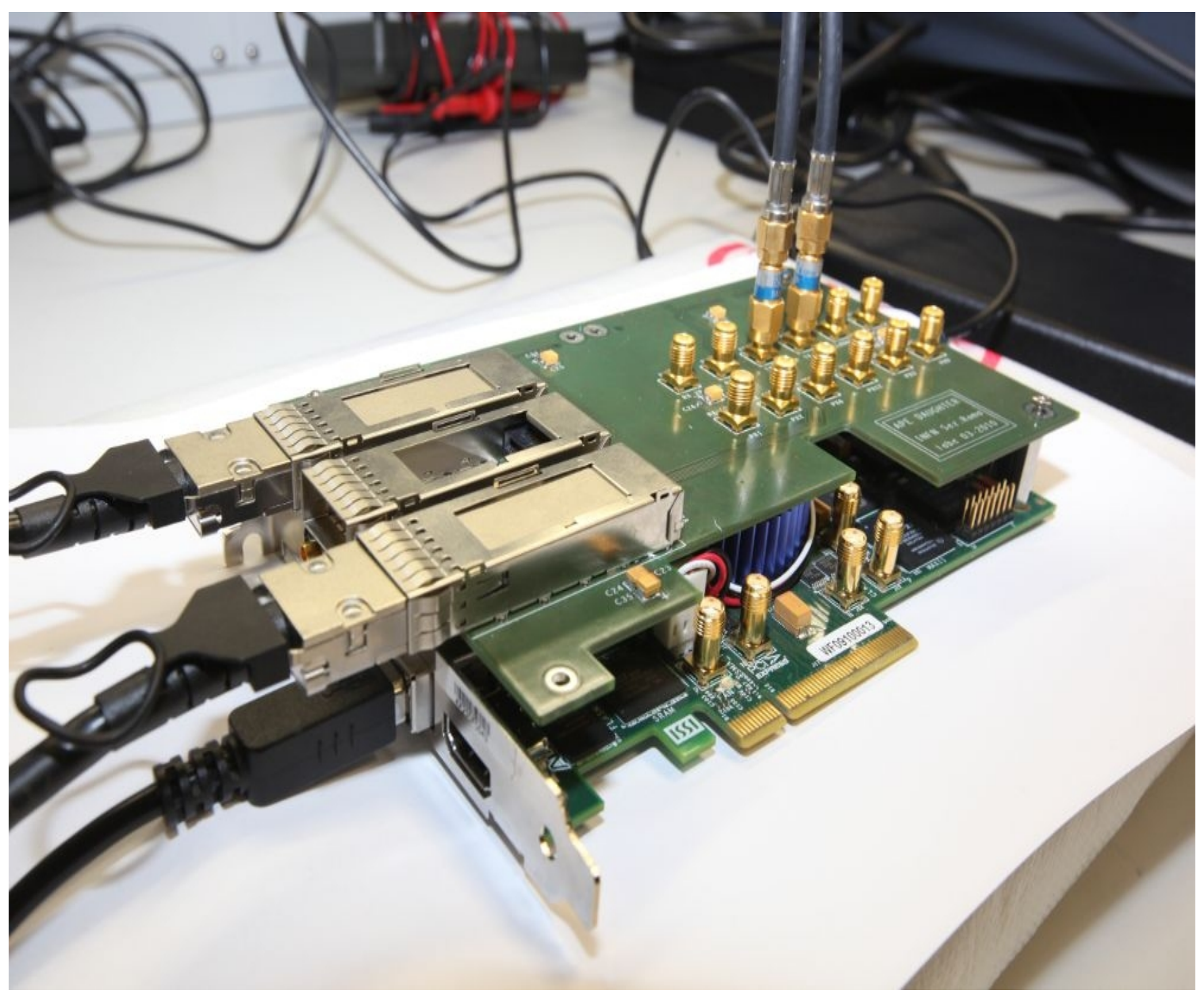

Figure 1: test system with Altera Stratix IV GX 230 development kit and custom mezzanine

Acknowledgments

APEnet+'s research is partially supported by European project EURETILE
http://www.euretile.eu

2 APEnet+ highlights

- APEnet+ is a packet-based direct network of point-to-point links with 2D/3D toroidal topology.
- Packets have a fixed size envelope (header+footer) and are auto-routed to their final destinations according to wormhole dimension-ordered static routing, with dead-lock avoidance.
- Error detection is implemented via CRC at packet level.
- Basic RDMA capabilities, PUT and GET, are implemented at the firmware level.
- Fault-tolerance features (will be added from 2011).

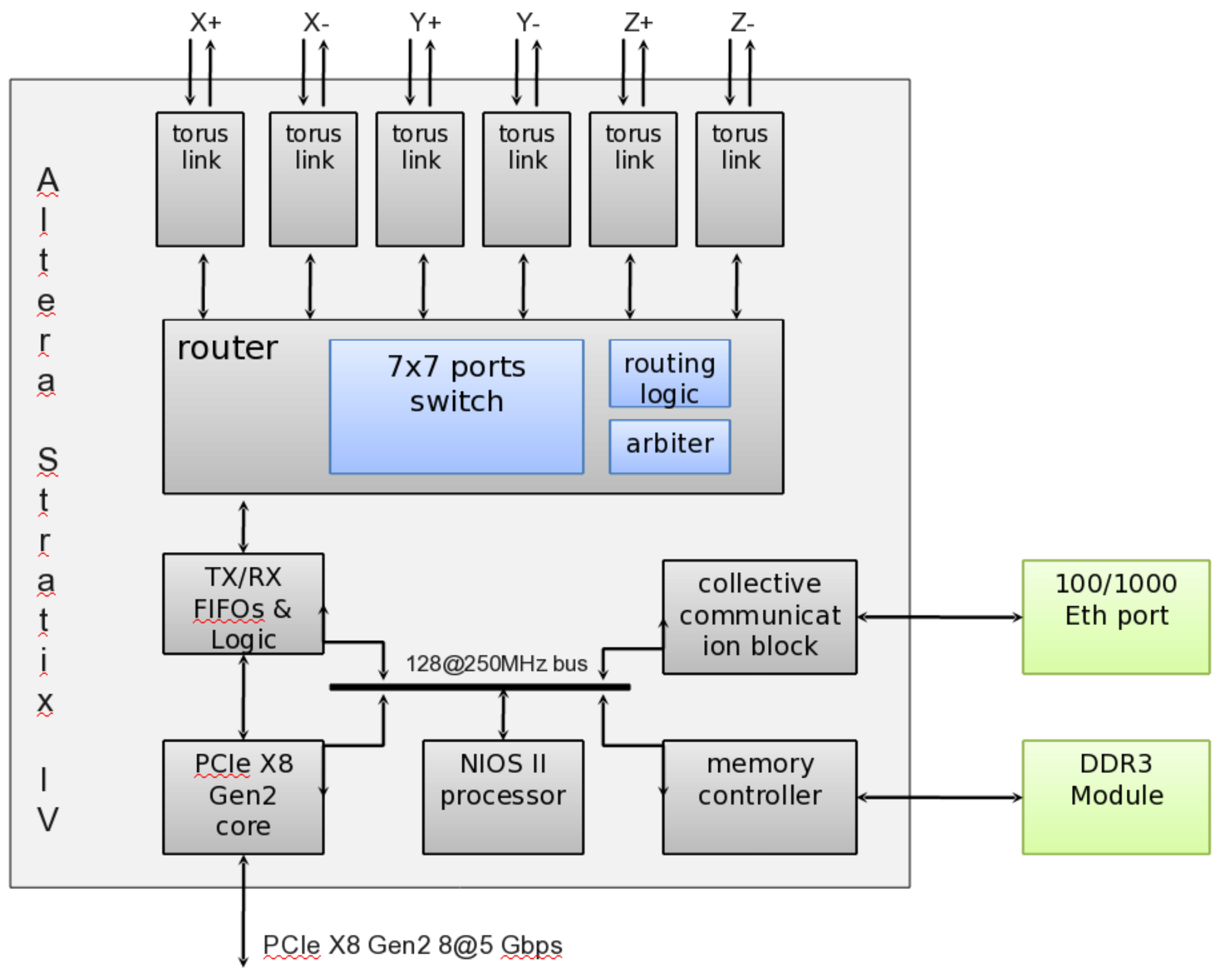

Figure 2: internal FPGA block architecture

3 Logic architecture

The majority of the logic modules are custom developed while the PCIe core is a commercial one.

High-level functionalities, like RDMA tables look-up, are carried on by a program running on an FPGA embedded processor (NIOS II), which uses the DDR3 module as both program- and data-memory.

The firmware block structure, depicted in figure 2, is split into a so called *network interface* (PCIe ,TX/RX logic, NIOS II processor, etc…) and a *router* (router component and torus links).

The router comprises a fully connected, 7-ports-in/7-ports-out switch, plus routing and arbitration blocks.

The routing block examines a packet header and resolves the destination address in a proper path across the switch. It supports the dimension- ordered routing algorithm, with a routing latency of 60ns.

Deadlock avoidance is implemented via the virtual channels technique, with 2 receive buffers on the torus link module.

Proper flow control is maintained via handshake of credits between a local RX block and the remote TX block, embedded in the link protocol data layer.

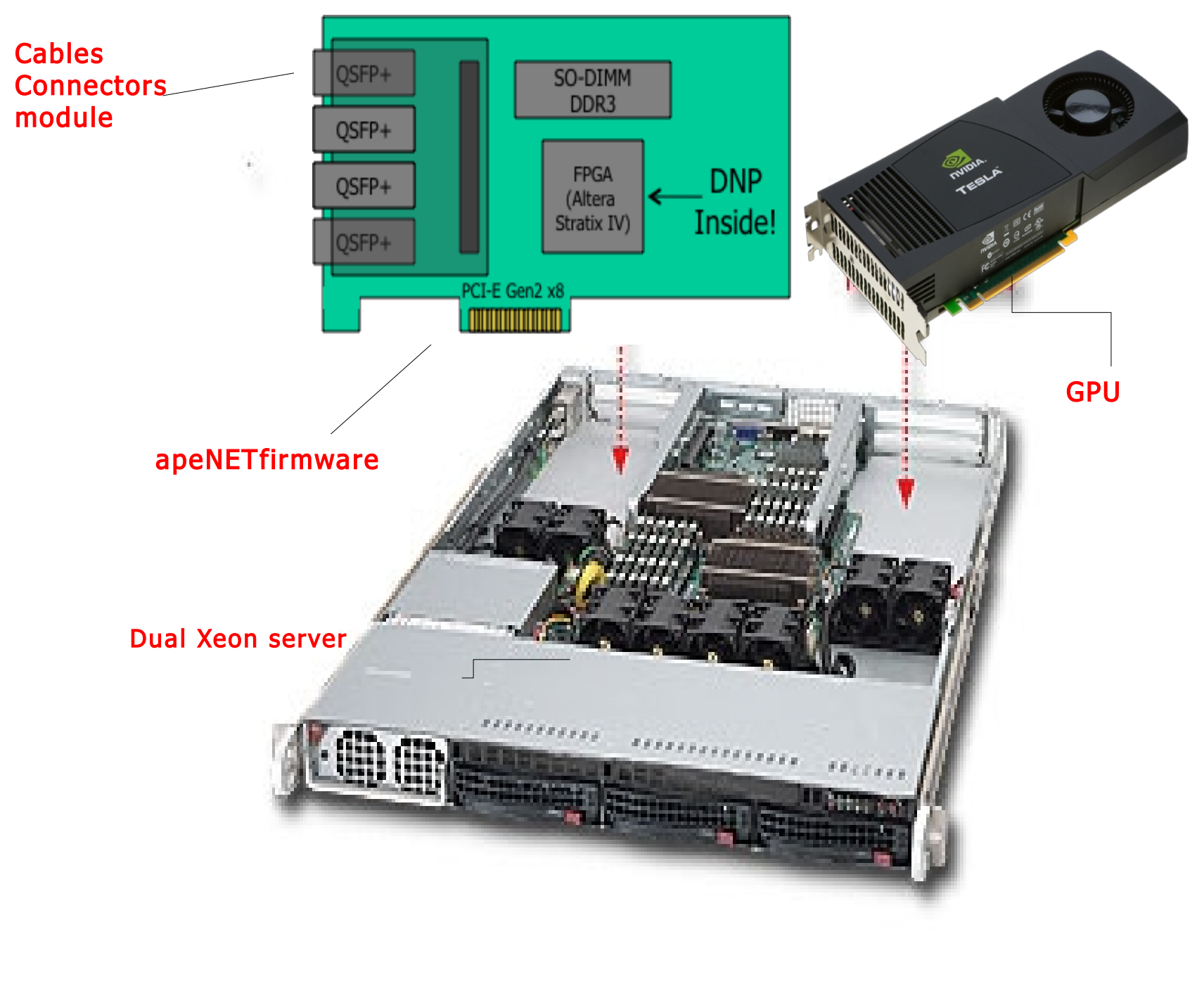

Figure 3: a sample U1 system with an APEnet+ card and a GPU

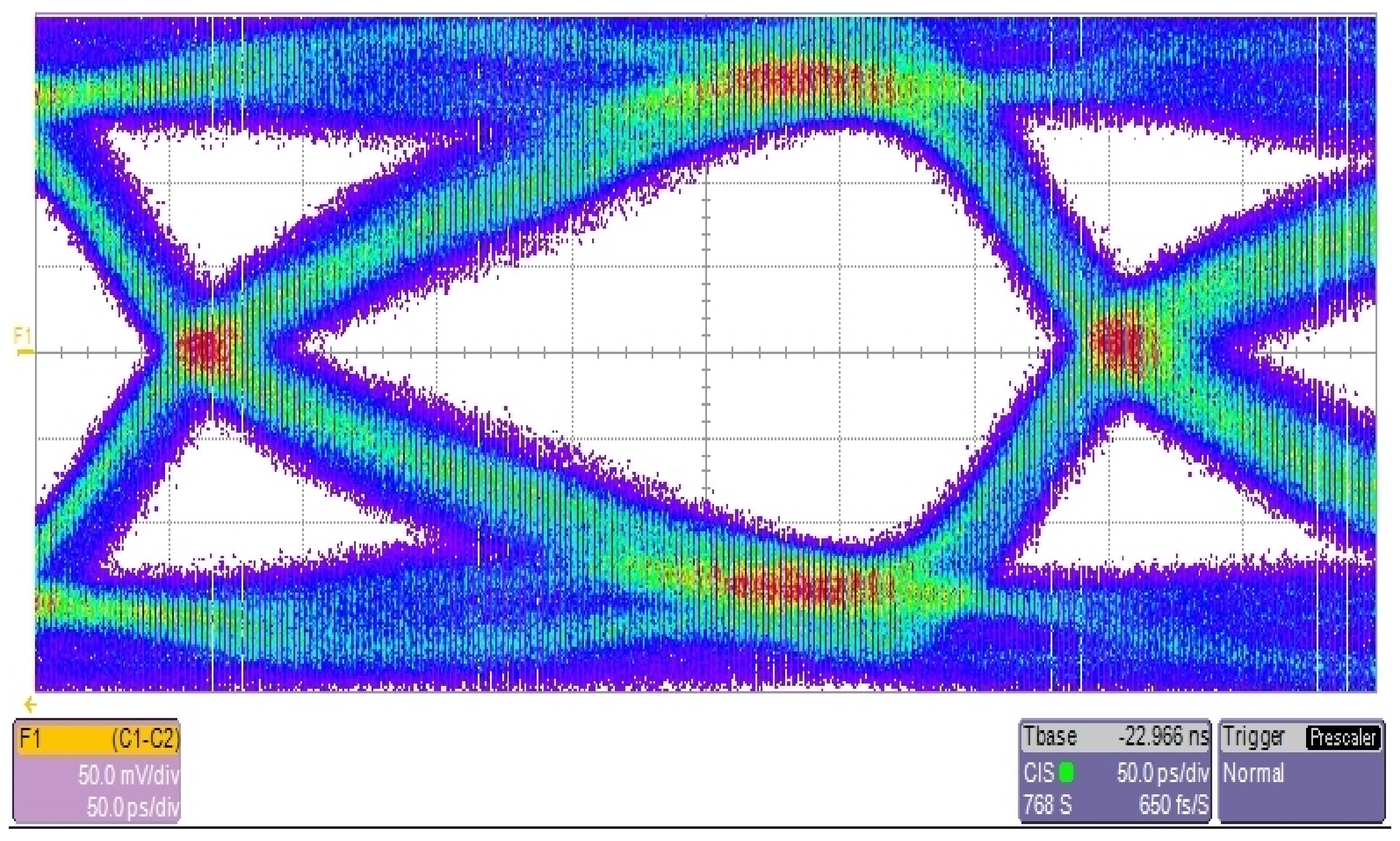

Figure 4: Eye diagram at 3 Gbps 2^{32} pseudo-random data stream

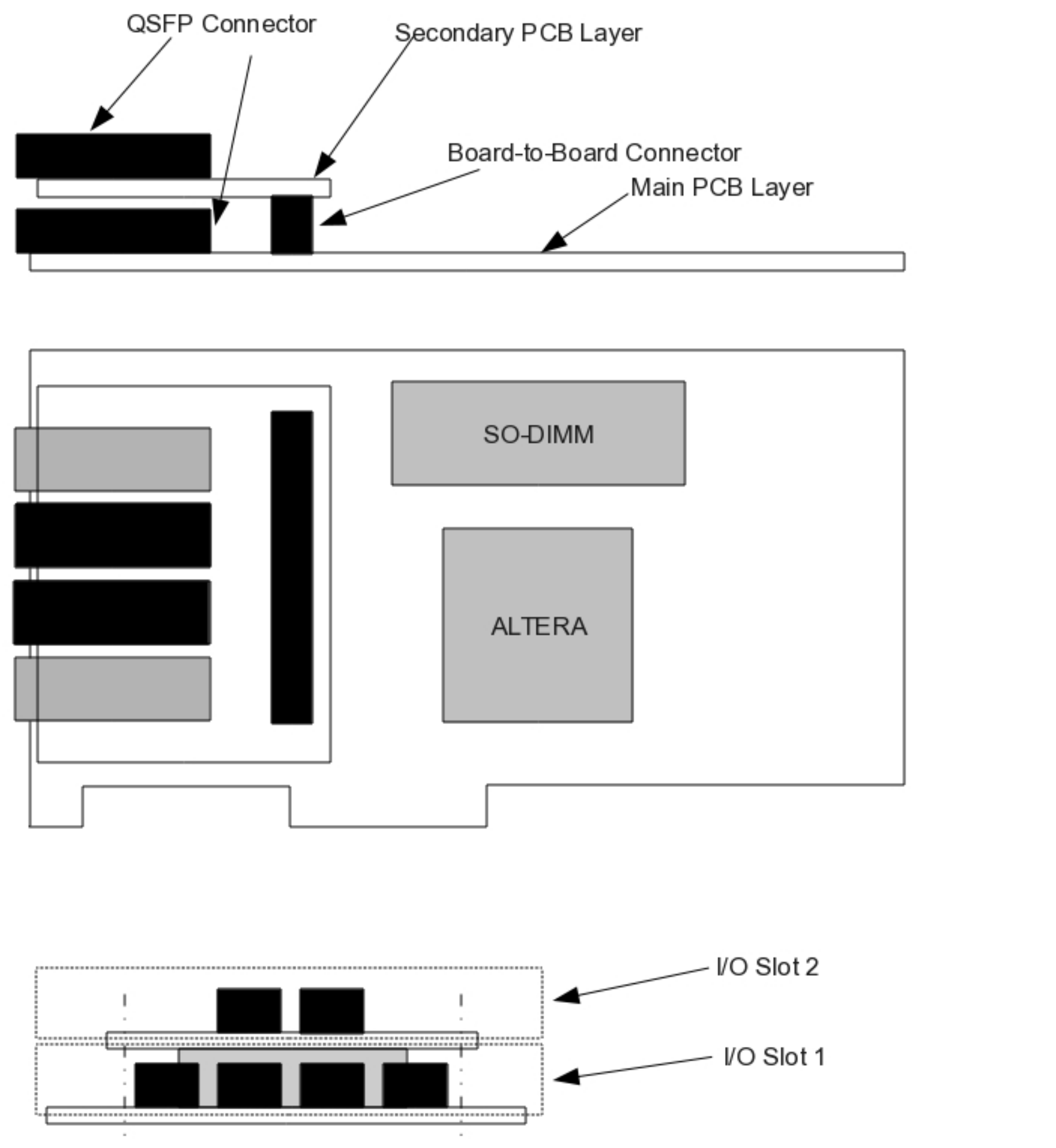

Figure 5: three views of the APEnet+ card, showing the secondary board which plugs on top of the main one to add two more links, thus enabling a 3D torus with 6 links total.

4 The hardware of the APEnet+ card

The APEnet+ hardware is a single FPGA-based PCI Express board, representing a vertex of a 3D torus mesh network (as depicted in figure 6 for a sample torus mesh of 4x2x2 sites) with 6 independent point-to-point multiple links channel (i.e. the links between mesh sites).

The on-board Altera Stratix IV FPGA (EP4SGX290) integrates 6 remote fully bidirectional channels, based on QSFP+ technology (up to 34 Gbps with 4 bonded FPGA transceivers) while the host connection is based on PCIe x8 v2.0 (4 GB/s peak bandwidth). A SO-DIMM DDR3 socket (512 MB – 2 GB) complements the board hardware resources.

The modular mechanics of the board allows to assemble the card in two different ways, depending on the cluster topology requirements:

- single slot width, 2D torus topology, (4 Torus Links);
- double slot width, 3D torus topology (6 Torus Links). Mechanics of the additional Torus Links are placed on a piggy-back module (figure 5).

The first prototype of APEnet+ card is expected for December 2010, while the FPGA firmware, the PCI Express interface and the physical layer interconnection technology have been developed and tested using an Altera development kit (Altera Stratix IV GX 230) and a daughter card (an HSMC mezzanine designed at LABE in INFN-Roma) hosting the final QSFP+ connectors for 3 Torus Links (figure 1).

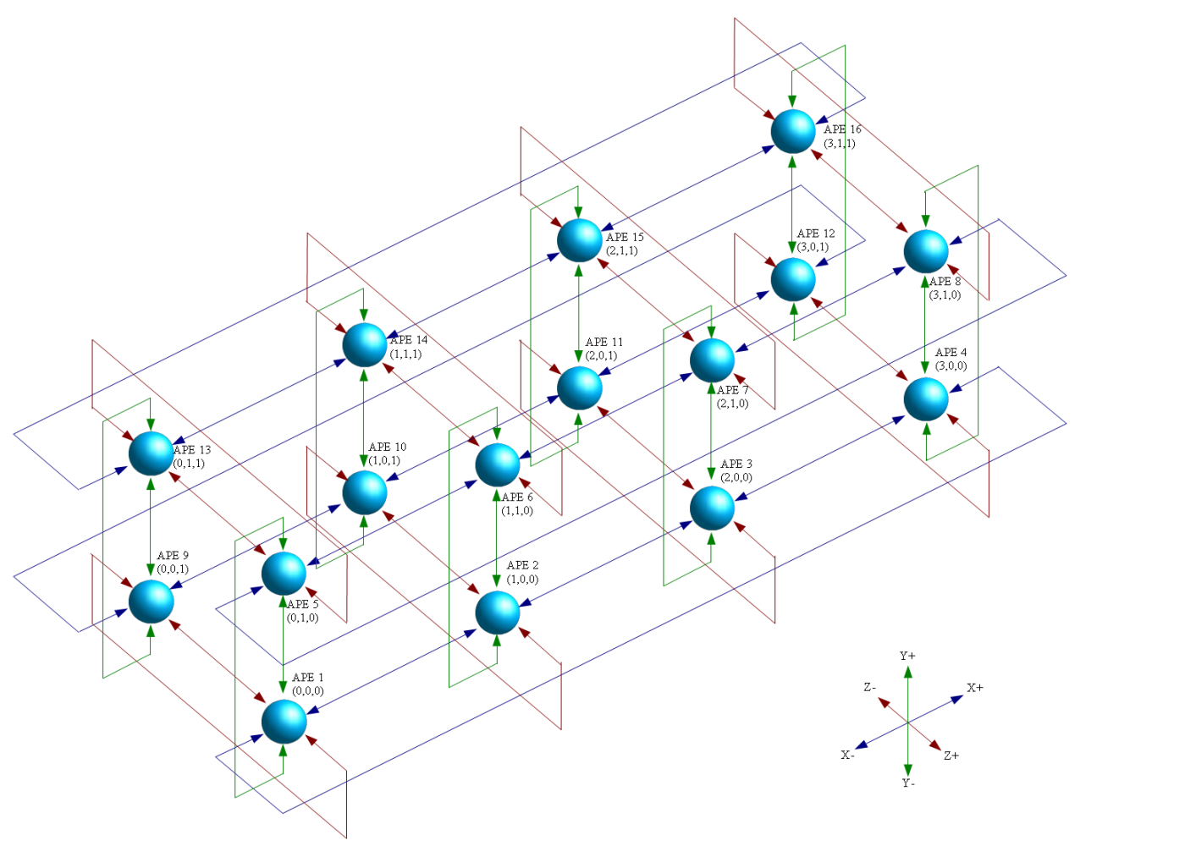

Figure 6: a pictorial representation of a 16 nodes cluster, arranged as a 4x2x2 torus.

5 APEnet+ programming

All APEnet+ software runs under Linux and is available under the GNU GPL Licence.

Of the two available sets of programming APIs, one is standard MPI and the other is a low-level custom RDMA one.

The RDMA APIs are available as a C language library:

- Communication primitives available to applications are: rmda_put(), rdma_get(), rdma_send().
- Memory buffer registration allows for exposing those buffers to RDMA primitives: register_buffer(), unregister_buffer().
- Events are routed to applications whenever RDMA primitives are executed by APEnet+: wait_event().
- The OpenMPI 1.X standard API is developed for APEnet+ as an adaptation BTL module, which is implemented atop the RDMA API.

6 Future work

We are currently exploring interconnection of GPU-equipped systems by means of APEnet+ (QUOnG project) to reach the PetaFLOPs range in aggregated computing power and working on some GPU-related driver optimizations.

For the 2011, our road-map foresees the integration of a “QUOnG rack”, a mesh of computing nodes which are rack-mounted 1U systems – based on a commodity Intel CPU Xeon 5650 – accelerated via high-end GPUs (Nvidia Tesla C1060/M2050) interconnected with the APEnet+ hardware. This system, housed in a single rack of 42U, will show a peak performance exceeding 56 TeraFlops and a power consumption of less than 26KW. Leveraging on APEnet+ network, multiple QUOnG racks can be assembled to push up the complete system to PetaFLOPs scale.

The presence on the APEnet+ card of a programmable component with a lot of free resources will allow us to explore **reconfigurable computing**, e.g. accelerating some tasks directly in hardware.

References

[1] First generation APENet is described in **arXiv:hep-lat/0409071**.
[2] For an overview of Lattice QCD, **arXiv:1002.4232v2** and references therein.
[3] APEnet+ web site is http://apegate.roma1.infn.it/APE